УДК 004.89:004.93

М.С. Клименко, Ф.В. Фомін
Інститут проблем штучного інтелекту МОН України і НАН України
пр. Академіка Глушкова, 40, м. Київ, Україна, 03680

РОЗРОБКА СТРУКТУРИ СИСТЕМИ РОЗПІЗНАВАННЯ ЕМОЦІЙНОГО СТАНУ ДИКТОРА

M.S. Klymenko, F.V. Fomin
Institute of artificial intelligence problems of MES and NAS of Ukraine
40, Academician Glushkov avenue, City of Kyiv, Ukraine 03680

DEVELOPMENT THE SYSTEM STRUCTURE FOR IDENTIFICATION SPEAKER EMOTIONAL CONDITION

У статті розглянуто сучасні підходи до автоматизованого розпізнавання емоцій і певних психологічних станів людини за її голосом. Запропоновано структуру системи ідентифікації емоцій, що використовує попередню обробку аудіо сигналу (шумозниження та сегментацію за учасниками), а також множини акустичних, просодичних та екстралінгвістичних характеристик мовлення для створення ознакового опису. Результати численнних досліджень вказують на необхідність застосування даних характеристик.
Ключові слова: модель диктора, акустичні характеристики емоцій, метод сумішей Гауса.

Modern approaches to automated recognition of emotions and psychological conditions by voice are described. The structure system for speaker emotion identification that uses a preprocessing audio signal (noise reduction and segmentation by participants) and a set of acoustic and prosodic features of speech and extra linguistic feature to create describing vector are proposed. The results of numeric research point to the necessity to use these characteristics together.
Keywords: speaker m odel, acoustic characteristics of emotions, the Gaussian mixture method.

Вступ

Автоматизоване розпізнавання емоційних станів на сьогодні є невирішеною проблемою. Водночас, застосування робастних методів ідентифікації емоційних станів дозволить зробити крок уперед у розробці людино-машинних інтерфейсів і систем контролю безпеки, що аналізують комплекс характеристик людської поведінки. Складнощі виникають внаслідок того, що людські емоції зазвичай зовні слабо виражені й швидко змінюються. Прояв емоцій людини може бути зафіксований зняттям показів датчиків фізичного стану (тиску, температури поверхні тіла та органів, електромагнітної активності мозку), але переважна більшість таких характеристик можуть бути отримані у безпосередньому контакті з людиною, що робить неможливим застосування характеристик на практиці. Віддалене розпізнавання можливо виконати візуально (рухи, міміка) та аудіально (за змінами у голосі).

Метою даного дослідження є розробка структури системи, здатної розпізнавати наявність певної емоції (із визначеної множини) у людини за її голосом. Застосування такої системи може бути використане у контролі за психічним станом пацієнтів під час діагностики або реабілітації. Аудіосигнал до системи може надходити із суттєвим рівнем шуму побутового характеру (акустичні викривлення приміщення, фоновий шум приладів, голоси співрозмовників). Ці обставини ускладнюють розробку системи, оскільки постає необхідність майже безперервного моніторингу проявів емоцій для негайного втручання спеціаліста або подальшого аналізу поведінки, але цим вони відрізняють її від існуючих наразі систем розпізнавання емоцій.

Сучасні підходи до розпізнавання емоційного стану за голосом

Типова схема розпізнавання емоцій може бути розділена на три етапи (вона представлена на рисунку 1). Після виокремлення фрагментів аудіосигналу із голосом виконується формування вектору ознак для подальшого порівняння з еталонними значеннями.

Для розпізнавання окремих емоцій краще за інші себе зарекомендували просодичні характеристики та їх комбінації [1, 2]. Якість розпізнавання тут повністю залежить від точності методів просодичного аналізу, автоматизувати які можливо тільки для конкретної мови або групи мов. Найкращою реалізацією автоматизованого розпізнавання емоцій за просодичними характеристиками є комп'ютерний детектор емоцій за голосом Voice-Stress Analysis, який розпізнає стресовий стан із ймовірністю 96%. Система є закритою й знаходить застосування в державних і правоохоронних органах США. До просодичних характеристик емоційних станів включають інтонаційний рисунок речень, інтенсивність мовлення, висоту та силу голосу (середню або ключових фрагментів речень).

У ряді робіт пропонуються характеристики мовлення диктора інших груп: акустичні й екстралінгвістичні [3, 4]. Ці характеристики мають менший спектр характеристик та нижчу робастність, але дозволяють робити оцінку зі значно меншим об’ємом обчислень. До акустичних характеристик емоційних станів відношення формант у голосних звуках, тривалість вимови фонем та пауз. Екстралінгвістичними ознаками є наявність специфічних подій, таких як зітхання, плач, сміх, кашель та ін. Такі ознаки дуже чітко характеризують певні множини емоцій, але наявність цих ознак не є обов’язковою, тому вони виступають допоміжними характеристиками.

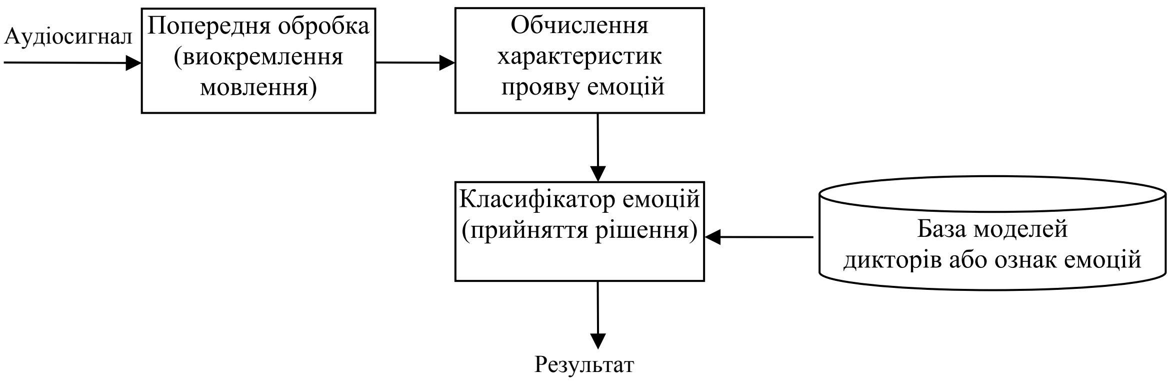


Рис. 1. Узагальнена типова схема розпізнавання емоцій за голосом

Завершальним етапом розпізнавання є класифікація отриманого вектору ознак із наявними еталонними значеннями для кожного диктора через високу варіативність ознак прояву емоцій. База еталонів може зберігати різні види інформації.

1. *Характеристики голосу диктора у стані спокою*. У цьому випадку тестовий вектор ознак порівнюється із еталонним для обраного диктора і, в разі відхилення, що перевищує поріг, робиться висновок про наявність певного емоційного прояву. Очевидно, що ймовірність правильно класифікувати прояви декількох емоцій низька через брак інформації про їх характеристики. Але таким чином зручно виконувати

перевірку знаходження диктора у стані підвищеної нервової напруженості [3].

2. *Характеристики голосу диктора у різних емоційних станах*. За такої інформації класифікатору достатньо почергово порівняти тестовий вектор ознак із усіма еталонними. У разі їх достатньої близькості, у просторі ознак робиться висновок про наявність певної емоції у диктора. Таким чином досягається найбільша точність розпізнавання. Недоліком такого підходу є необхідність створювати множину моделей для кожної емоції диктора, яку потрібно розпізнавати, у той час, коли для навчання системи може не існувати запису голосу диктора у необхідному емоційному стані.

3. *Загальні характеристики різних емоційних станів*. Індивідуальні характеристики диктора не зберігаються зовсім через узагальненість еталонних моделей, які створюються при навчанні системи на великій вибірці дикторів. У цьому випадку рішення про наявність певної емоції у диктора приймається, коли тестовий вектор ознак близький до відповідного еталонного вектору ознак емоції. Системою зберігається найменший обсяг інформації, що відображається на якості роботи даного підходу: через варіативність індивідуальних проявів емоцій узагальнені вектори ознак втрачають свої якості, наближаючись один до одного і збільшуючи помилку хибної ідентифікації.

Як класифікатор для даної задачі застосовується метод, який кращим чином може створити вирішальне правило на обраній множині характеристик. Так, для просодичних характеристик, де є дані різної структури (графіки, числові послідовності), застосовують нейромережі, приховані марківські моделі. Для більш одноманітних векторів ознак, акустичних (спектральних) та екстралінгвістичних, можливе використання простіших методів лінійного квантування, сумішей Гауса і т.п.

Сучасні розробники уникають розпізнавання окремих емоцій, фокусуючись на детекції простіших психологічних станів (афекту, стресу, тонусу), які включають у себе прояв одразу низки емоцій. У даній статті пропонується розпізнавання базових емоцій із можливістю подальшого розширення їхнього переліку.

Постановка задачі

Виходячи з мети даної роботи, поставлені наступні задачі:

1. На основі існуючих досліджень відібрати перелік акустичних ознак, що дозволяють охарактеризувати якомога більшу кількість емоцій людини.
2. Скласти перелік основних емоцій для розробки системи та чисельних досліджень.
3. Розробити структуру системи, що здатна розпізнавати емоції диктора з обраної множини по аудіосигналах, записаних у побутових умовах.
4. Дослідити внесок акустичних ознак у розпізнавання емоцій.

Опис системи розпізнавання

Існує чимало класифікацій людських емоцій, не будемо зупинятись на конкретних, що налічують від 20 і більше найменувань через складність розпізнавання за голосом схожих проявів різних емоцій. Прийнято виділяти базові емоції, яких є декілька класифікацій, наприклад, загальновідомою є класифікація К.Ізарда з 11 станів [6]. У даній роботі ця множина звужена до 5 базових емоцій, які частіше за все обираються у гештальт-психології і мають більшу, відносно інших емоцій, амплітуду зміни поведінки як за голосом, так і невербальними проявами [7]. Отже, розроблювана система повинна розпізнавати наступні емоції: радість, інтерес, страх, сум, злість.

Структура системи розпізнавання емоцій, згідно з постановкою задачі, представлена на рисунку 2 і різниться від типової появою додаткових етапів попередньої обробки, а також баз даних ідентифікаційних моделей дикторів та емоційних станів.

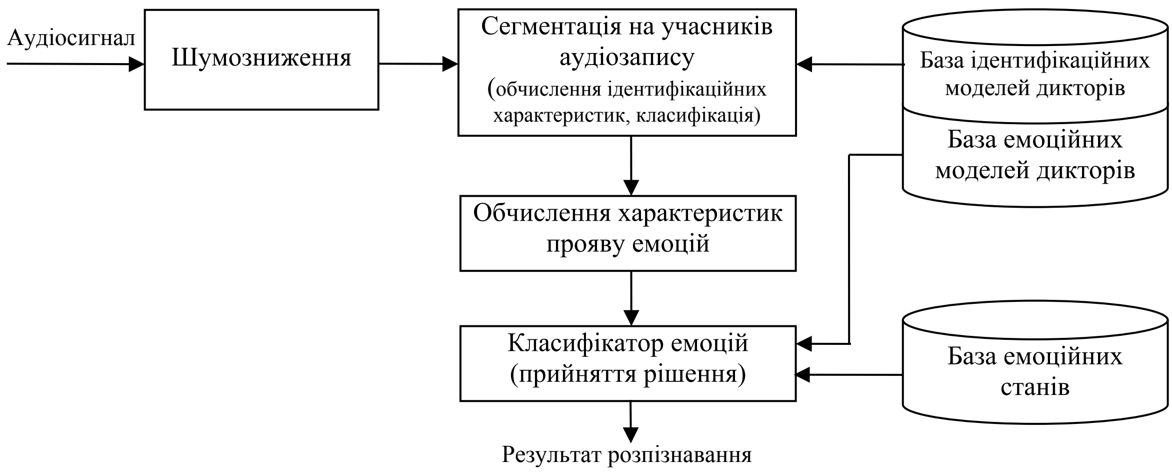


Рис. 2. Структурна схема системи розпізнавання емоцій диктора за характеристиками голосу

Шумозниження. У кожному звуковому фрагменті виконується розпізнавання наявності та зменшення амплітуди адитивних стаціонарних шумів за допомогою спектрального віднімання смуг, у яких не фіксується людський голос.

$$Y(f,t) = \max\left\{1 - k\frac{W(f,t)}{X(f,t)}, 0\right\} \cdot X(f,t),$$

де f – частота сигналу,
t – початок відрізку сигналу за часом або відліками,
$X(f,t)$ – амплітудний спектр сигналу,
$W(f,t)$ – амплітудний спектр шуму,
$Y(f,t)$ – амплітудний спектр очищеного сигналу,
k – коефіцієнт зниження шуму.

Сегментація на учасників. Для виконання цієї операції в даній роботі використана підсистема ідентифікації дикторів із доробку авторів [5], яка сегментує увесь сигнал як на відрізки, що належать одному із учасників сигналу, так і на відрізки пауз (тобто фонового шуму). Слід зауважити, що більшість екстралінгвістичних характеристик на цьому етапі будуть віднесені до пауз без розпізнавання належності до учасника розмови. Оскільки метод ідентифікації диктора використовує моделі ознак за окремими широкими фонетичними класами (множинами споріднених за акустичними характеристиками фонем), то в результаті цього етапу отримуємо одночасно ще одну сегментацію – на фонетичні класи, інформація про які використовується при обчисленні акустичних характеристик.

$$Seg(X_i) = \{F(X_i), Sp(X_i), N(X_i)\}$$

де X_i – відрізок сигналу, отриманого на етапі шумозниження,
$F(X_i)$, $Sp(X_i)$, $N(X_i)$ – перелік відрізків окремих фонем.

Сегментація на учасників є текстонезалежною, тобто вона не враховує повну інформацію мовлення, а лише дані про фонеми. Це не дає змогу отримати ряд просодичних характеристик без залучення додаткових методів, тому, наразі,

розмітка на речення та смислові конструкції виконується оператором.

На етапі *обчислення характеристик прояву емоцій* за наборами відрізків цільового диктора отримуються вектори ознак, які будуть класифіковані на наступному етапі. Перед формуванням вектора ознак емоцій та способу їх обчислення необхідно обрати множину емоцій, під які буде налаштована система. Простір ознак для даних емоцій сформовано із 7 акустичних, просодичних та екстралінгвістичних характеристик. Згідно з порівняльним аналізом досліджень [6, 7] зміни вербальних характеристик для певних емоційних станів наведено у таблиці 1. Таким чином створено простір ознак, у якому можливо виокремити особливості прояву емоцій за голосом. Відсоткові значення відхилень були отримані усередненням персональних вимірів за тестовою множиною дикторів при навчанні системи.

Таблиця 1. Порівняльний аналіз зміни вербальних характеристик для певних емоційних станів людини відносно спокійного стану

Характеристики	***Емоції***				
	Радість	*Інтерес*	*Страх*	*Сум*	*Злість*
Висота голосу, Гц	Значно підвищена (від 11%)	Підвищена (від 7%)	Значно підвищена (від 13%)	Понижена (-8% й нижче)	
Сила голосу, дБ		Підвищена (від 16%)	Підвищена (від 7%)	Понижена (-6% й нижче)	Підвищена (від 9%)
Відстань між формантами у голосних звуках, Гц	Збільшення відстані між *F2* та *F3* (на 5-8%)	Збільшення відстані між *F2* та *F3*, особливо для звуків {О, А} (до 6%)	Зменшення відстані між *F2* та *F3* (-6% й нижче)	Зменшення відстані між *F1* та *F2* (від -4%)	Збільшення відстані між *F2* та *F3* (від 5%)
Тривалість вимови складів, мс	Прискорена (від 6%)	Прискорена (від 9%)			
Тривалість пауз, мс				Подовжена (від 5%)	Зменшена (від -6%)
Наявність кашлю, зітхань, плачу, сміху	Можливі сміх, плач	Можливі зітхання	Можливий плач	Можливі зітхання, плач	Можливі кашель, плач
Зміна інтонації у реченні, графік Гц від часу	Без змін	Підвищення до кінця	Без змін	Підвищення на початку	Підвищення на початку

Для отримання акустичних характеристик дикторів використані згладжені спектри широких фонетичних класів, за якими легко розрахувати висоту голосу (основний тон), силу голосу та відстань між формантами у голосних звуках. Характеристики тривалості вимови складів та пауз між словами обчислюються усереднено по реченню, а зміна інтонації у реченні є динамічною просодичною характеристикою. Для спрощення задачі класифікації інтонаційних особливостей різних мов, часова характеристика інтонації (висота тону) сегментується на наступні ділянки для кожного речення:

- підвищення тону до кінця речення;
- підвищення тону на початку речення;

- спад тону усередині речення;
- підвищення тону усередині речення.

Уточнення кривої висоти тону та порівняння інтонаційних векторів тільки ускладнює вектор ознак і робить моделі дикторів сильно залежними від особливостей мови або діалекту.

Екстралінгвістичними характеристиками є наявність у фрагментах пауз (шуму) однієї з акустичних подій, притаманних обраним емоціям: кашлю, зітхань, плачу, сміху. Для автоматизації пошуку даних подій у блоку сегментації до моделі загального шуму були створені додаткові моделі екстралінгвістичних подій. Однак, складність створення універсальних моделей у цьому випадку наразі змушує використовувати ручну розмітку, а неможливість системи ідентифікувати диктора вирішується припущенням щодо його належності учаснику розмови, фрагмент мовлення якого знаходиться найближче за часом до даного фрагменту.

На етапі *прийняття рішення* використовується сформований на попередньому етапі набір векторів ознак, який оцінюється одним із двох вирішальних правил:

$$R1(fv, Sp, e) = \sum_{i=1}^{7} a_i \cdot C_i(fv, DB_e, Sp_i), \quad (1)$$

де fv – набір векторів ознак тестового сигналу,
Sp – модель диктора (значення 7 характеристик голосу у стані спокою),
DB_e – модель еталонної емоції,
e – номер еталонної емоції із переліку обраних до розпізнавання,
C_i – простий класифікатор за i-ю характеристикою,
a_i – ваговий коефіцієнт.

або

$$R2(fv, Sp) = \arg\max_{e=1..5} (R1(fv, Sp, e)) \geq p, \quad (2)$$

де p – порогове значення.

Простим класифікатором $C_i(fv, DB_e, Sp_i)$ у даному підході виступає метод сумішей Гауса, який виконує оцінку приналежності набору векторів ознак fv до еталонної моделі DB_e моделі диктора Sp. Вирішальне правило (1) базується на методі *AdaBoost*, використаного раніше у системі ідентифікації диктора [5], і здійснює зважування внесків простих класифікаторів для мінімізації похибки розпізнавання. Значення, що отримуються вирішальним правилом (1), показують відповідність тестової вибірки емоції з номером e моделі диктора Sp. Значення на виході бінарного класифікатора за методом *AdaBoost,* наближаються до нуля за відсутності впевненості у результаті розпізнавання. В інших випадках *R1* може дорівнювати максимальному значенню приналежності тестової вибірки до еталонних моделей.

Модель емоції DB_e складається із 3 обмежувальних параметрів за кожною із характеристик голосу:

- тип обмеження (немає, зверху, знизу, діапазон);
- значення обмеження (число або 2 числа діапазону);
- абсолютна (в одиницях виміру характеристики за таблицею 1) чи відносна (у відсотках) шкала.

Вирішальне правило (1) використовується для визначення прояву емоції у аудіофрагменті. Натомість вирішальне правило (2) за пороговим значенням

визначає, яка з емоцій була проявлена.

У структуру було додано базу даних емоційних моделей дикторів. Вона різниться від аналогічної бази для сегментації по учасниках розмови низкою ознак. По-перше, вектор ознак для сегментації формується із мел-честотних кепстральних коефіцієнтів, що є стійкими до низки емоційних проявів, а отже, не є інформативними у даній задачі. Для бази емоційних моделей використано набір вище зазначених характеристик, які отримані з аудіофрагменту мовлення диктора у спокійному стані (тобто без прояву емоцій, що підлягають розпізнаванню). Емоційна модель диктора *Sp* представляє собою дані суміші Гауса, отримані за 7 характеристиками голосу.

Відсутність необхідності швидкого пошуку по базі емоційних станів (диктор вже відомий із попереднього етапу) спрощує реалізацію. У базі сегментації для цього створено ієрархічну пошукову структуру, де схожі за ознаками моделі дикторів згруповані між собою. Окрім того, пошук по базі емоційних станів взагалі не використовується, вона наразі є реферативною, а додаткова інформація, яку вона містить, може додатково використовуватись при сегментації учасників розмови, оскільки сукупність характеристик містить індивідуальні особливості дикторів.

База емоційних станів представляє собою набір дикторонезалежних записів щодо діапазону відхилень узагальненого вектору ознак за прояв певної емоції від узагальненого вектору ознак стану спокою. Саме поєднання цієї інформації із даними про особливості голосу диктора дає змогу встановити індивідуальний діапазон зміни вектору ознак. Інформація по діапазонах характеристик зберігається у відносному вигляді.

Для чисельного дослідження ефективності використання запропонованої структури системи із використанням додаткових емоційних моделей брали участь 10 дикторів з різними голосовими даними (5 жінок і 5 чоловіків віком від 14 до 68 років, m=34, sd=11). Для побудови моделей були записані фрагменти емоційно забарвленої мови дикторів загальною середньою тривалістю 1 хвилина. Запис здійснювався динамічним мікрофоном у приміщенні зі слабкими сторонніми шумами (рівень шуму 25dB) з частотою дискретизації 44,1 кГц і глибиною квантування 16 біт.

Для проведення порівняльного аналізу був реалізований метод [5] із використанням бустінг-алгоритму AdaBoost, що дозволило зробити порівняльний аналіз внеску характеристик у розпізнавання емоцій за різної довжини тестового аудіофрагменту, наведений на рисунку 3. По вертикальній осі наведено значення нормованих бустінг-коефіцієнтів, отриманих після навчання на тестових зразках.

Із діаграми видно, що за навчання на короткочасних фрагментах внески характеристик виявляються приблизно рівними. А за наявності великої кількості інформації найвагомішими стають показники міжформантних відстаней та інтонаційного малюнку, трохи менш вагомими є висота голосу та екстралінгвістичні події. Інші характеристики носять другорядний характер.

Що стосується ймовірності розпізнавання, то на моделях, побудованих на вимові одного речення (до 15 секунд), вона дорівнює в середньому 64% ± 3,1% ($p < 0,05$), при збільшені об'єму навчальної інформації зростає до 77% ± 3,6% ($p < 0,05$), а за 60-секундних навчальних зразків дорівнює 82% ± 2,7% ($p < 0,05$).

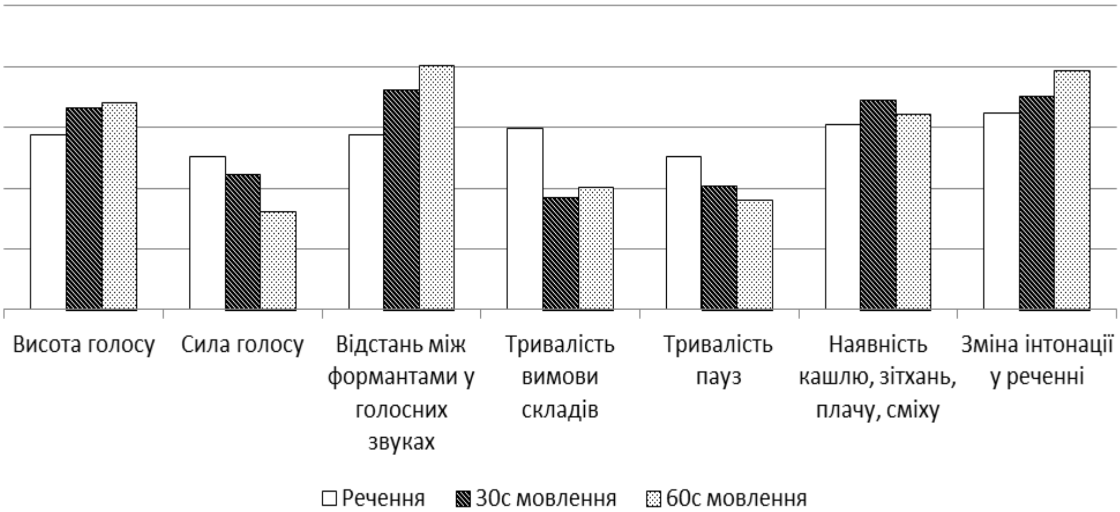


Рис. 3. Порівняльна діаграма внеску характеристик у розпізнавання емоцій за різної довжини тестового аудіофрагменту

Висновки

У даній статті виконано огляд сучасних підходів до розпізнавання емоцій за голосом, розроблено структуру системи такого розпізнавання, в якій використовуються моделі емоцій та дикторів. Аналіз отриманих результатів дозволив зробити наступні висновки.

1. Існуючі методи розпізнавання емоцій за голосом у високій мірі чутливі до якості передачі мовленнєвого сигналу та не мають змоги пристосовуватись до особливостей вимови диктора.

2. Для чисельних досліджень обрано 5 базових емоцій, які мають більшу відносно інших амплітуду зміни поведінки вербальної активності. Для їх векторного опису відібрано множину із 7 акустичних, просодичних та екстралінгвістичних характеристик, що дозволяють розширити простір ознак, у якому можна виокремити особливості прояву емоцій за голосом.

3. Запропоновано підхід до проектування системи розпізнавання емоцій за голосом, який використовує блоки шумозниження, сегментації на учасників розмови та паузи як попередню обробку. Це дозволило отримати додаткову інформацію щодо пофонемної розмітки, яка використовується при обчисленні акустичних характеристик. Крім того, запропоновано зберігання окремих емоційних моделей дикторів та узагальнених моделей емоцій для можливості пристосування до диктора.

4. Проведено чисельне дослідження якості автоматичного розпізнавання та класифікації за усією множиною характеристик. Середня ймовірність розпізнавання емоцій за навчання моделей на 60-секундних фрагментах 10 дикторів сягає 82%. Також, досліджено внесок окремих характеристик у вирішальне правило, який показав, що найвагомішими є показники міжформантних відстаней та інтонаційного рисунку.

5. Складність для класифікації представляють ділянки, що містять екстралінгвістичні фрагменти (через неможливість системи ідентифікувати диктора таких голосових дій) та просодичні характеристики (через ручну розмітку аудіоматеріалу). Вирішення цих проблем може стати подальшим розвитком даної роботи.

Література

1. Voice Stress Analysis Services // [Електр. Ресурс]. - Режим доступу: http://www.voicestressanalysis.net
2. Потапова Р. К. О возможности перцептивно-слухового распознавания состояния «агрессия» по устной речи / Потапова Р. К., Комалова Л. Р. // Вестник МГЛУ. – 2014. – №13 (699) . – С.202-214.
3. Brose A. Affective states contribute to trait reports of affective well-being / Brose A., Lindenberg U., Schmiedek F. // Emition .- 2013(5) .- p. 940-948.
4. Kächele M. Prosodic, Spectral and Voice Quality Feature Selection Using a Long-Term Stopping Criterion for Audio-Based Emotion Recognition / M. Kächele, D. Zharkov, S. Meudt and F. Schwenker // Pattern Recognition (ICPR), 2014 22nd International Conference on, Stockholm . – 2014. – pp. 803-808.
5. Клименко Н. С. Сегментация и классификация участков речевого сигнала для построения моделей дикторов в системе идентификации говорящего//Мат. Междунар. научно-технич. конф. "Искусственный интеллект. Интеллектуальные системы". - Донецк: ІПШІ "Наука і освіта", 2012. - С. 80-82.
6. Психология эмоций / Изард К.Э. Перев. с англ. – СПб: Издательство «Питер», 1999. – 464 с.
7. Черняев Л.С. Взгляд гештальт-терапевта на базовые эмоции / Л.С. Черняев // Гештальт 2007. – Ч 2. – Философия и этика в гештальт-подходе. – М.: МГИ, 2007. – С.17-24.

References

1. Voice Stress Analysis Services // [El. recource]. – Access mode: http://www.voicestressanalysis.net
2. Potapova R. K. O vozmozhnosti pertseptivno-sluhovogo raspoznavaniya sostoyaniya «agressiya» po ustnoy rechi / Potapova R. K., Komalova L. R. // Vestnik MGLU. – 2014. – #13 (699) . – S.202-214.
3. Brose A. Affective states contribute to trait reports of affective well-being / Brose A., Lindenberg U., Schmiedek F. // Emition .- 2013(5) .- p. 940-948.
4. Kächele M. Prosodic, Spectral and Voice Quality Feature Selection Using a Long-Term Stopping Criterion for Audio-Based Emotion Recognition / M. Kächele, D. Zharkov, S. Meudt and F. Schwenker // Pattern Recognition (ICPR), 2014 22nd International Conference on, Stockholm . – 2014. – pp. 803-808.
5. Klimenko N. S. Segmentatsiya i klassifikatsiya uchastkov rechevogo signala dlya postroeniya modeley diktorov v sisteme identifikatsii govoryaschego // Materialyi Mezhdunarodnoy nauchno-tehnich. konf. "Iskusstvennyiy intellekt. Intellektualnyie sistemyi". - Donetsk: IPShI "Nauka I osvIta", 2012. - S. 80-82.
6. Psihologiya emotsiy / Izard K.E. Perev. s angl. – SPb: Izdatelstvo «Piter», 1999. – 464 s.
7. Chernyaev L.S. Vzglyad geshtalt-terapevta na bazovyie emotsii / L.S. Chernyaev // Geshtalt 2007. – Ch 2. – Filosofiya i etika v geshtalt-podhode. – M.: MGI, 2007. – S.17-24.

RESUME

M.S. Klymenko, F.V. Fomin

Development the system structure for identification emotional speaker condition

Modern approaches to automated recognition of emotions and psychological conditions by voice are described in he article. The stages of typical emotion recognition system are described. It allows to evaluate the current state of solving the problem. The article shows the different types of information for modelling purposes and analysis of the advantages and disadvantages of each type.

The proposed structure of emotion identification system uses a preprocessing audio signal stage (it includes noise reduction and segmentation by participants) and a set of acoustic, prosodic and extralinguistic features of speech to create feature vector. Database of emotional states is a speaker-independent set of records on a range of deviations between generalized feature vector of certain emotions and the generalized feature vector at normal condition. Instead, speaker database keeps individual characteristics of pronunciation. Combining this information with data on the general characteristics of emotion lets us set individual range of feature vector variation.

The results numeric researches of the 5 basic emotions showed the likelihood of automatic recognition of emotional states at 82%. The characteristics of formant distances and intonation figure used to contribute most to the decision rule among all characteristics. It points to the necessity for these characteristics to use together.

Speaker identification on fragments of extralinguistic events and automation of prosodic features computing may increase the likelihood of emotional states recognition and become a further development of this system.

М.С. Клименко, Ф.В. Фомін
Розробка структури системи розпізнавання емоційного стану людини

У статті розглянуто підходи до автоматизованого розпізнавання емоцій та певних психологічних станів людини за її голосом. Наведено етапи типової схеми розпізнавання емоцій, що дає змогу оцінити сучасний стан вирішення задачі. Описано різні типи інформації для формування еталонів розпізнавання із аналізом переваг та недоліків кожного типу.

Запропоновано структуру системи ідентифікації емоцій, що має наступні вдосконалення: попередня обробка аудіосигналу (шумозниження та сегментацію за учасниками), використання множини акустичних, просодичних та екстралінгвістичних характеристик мовлення для створення ознакового опису емоційних станів і моделей дикторів. База емоційних станів представляє собою набір дикторонезалежних записів щодо діапазону відхилень узагальненого вектору ознак при прояві певної емоції від узагальненого вектору ознак у стані спокою. Натомість база моделей дикторів зберігає індивідуальні характеристики вимови. Поєднання цієї інформації із даними про загальні характеристики емоцій дає змогу встановити індивідуальний діапазон зміни вектору ознак.

Результати чисельних досліджень із 5 базовими емоціями показали ймовірність автоматичного розпізнавання емоційних станів на рівні 82%. Показники міжформантних відстаней та інтонаційного рисунку серед використаних характеристик роблять найбільший внесок у вирішальне правило класифікатора, що вказує на необхідність застосування даних характеристик.

Ідентифікація диктора на фрагментах екстралінгвістичних подій та автоматизація обчислення просодичних характеристик може підвищити ймовірність розпізнавання емоційних станів і стати подальшим розвитком системи.

Надійшла до редакції 06.09.2016

